Computing and Informatics, Vol. 45, 2026, 488–512, doi: 10.31577/cai_2026_2_488

MANIFOLD-AWARE HISTORICAL TOPOLOGY MAPPING FOR FOUNDATION MODEL-BASED ROBOTIC NAVIGATION

Haibo Li, Zesheng Zhan, Yaming Yang

Department of Computer Science and Technology, School of Informatics
Xiamen University, Xiamen 361000, China
&
Shenzhen Research Institute, Xiamen University
Shenzhen 518000, China

Fansheng Meng

Wenzhou Taiyi Intelligent Technology Co., Ltd.
Wenzhou 325000, China
&
Taichang Technology (Hangzhou) Co., Ltd.
Hangzhou 310000, China

Yaoxin Chen

Institute of Artificial Intelligence, Xiamen University
Xiamen 361000, China

Chong Zhao, Qicong Wang*

Department of Computer Science and Technology, School of Informatics
Xiamen University, Xiamen 361000, China
&
Shenzhen Research Institute, Xiamen University
Shenzhen 518000, China
e-mail: qcwang@xmu.edu.cn

* Corresponding author

Abstract. Vision-and-language navigation faces critical challenges, including insufficient long-horizon reasoning, misalignment between instruction segments and visual observations during navigation, and hallucination interference in large model reasoning. These limitations hinder the ability of large models to accurately interpret complex scenes through textual summaries alone, particularly in understanding spatial distributions of environmental elements and dynamically adapting to the demands of target instructions. To address these challenges, we propose a novel historical-topological vision-and-language navigation framework based on multi-curvature spatiotemporal Chain-of-Thought reasoning. Leveraging an encoder-decoder architecture, our model encodes historical information from input features, effectively resolving the lack of global topological guidance in long-horizon reasoning tasks. To mitigate hallucination in large model reasoning, we exploit the inherent data characteristics of vision-and-language navigation tasks. By leveraging multi-curvature manifolds, we identify salient spatiotemporal references in visual observation sequences, enabling the model to ground its reasoning in these references for accurate situational understanding and future prediction. To enhance spatial structure perception, we introduce a graph self-attention mechanism that jointly models node distances and visual similarity in historical topological graph embeddings, effectively capturing spatial relationships between nodes. This work introduces novel insights for large model-based vision-and-language navigation in indoor scenes, advancing their applicability in complex environments.

Keywords: Multimodal large models, hyperbolic manifolds, landmark, chain of thought, navigation

1 INTRODUCTION

Recent studies [1, 2, 3, 4, 5] on vision-language navigation (VLN) using foundation models demonstrate that large language models (LLMs) can comprehensively describe an agent's movement trajectories and task progress during navigation. To equip these language models with visual perception capabilities, existing methods typically convert the agent's sequential environmental observations into textual descriptions via image captioning models, subsequently summarizing past observations and decisions into pure text formats for experience tracking and future planning. The LLM then integrates this information to infer actions that guide the agent toward its goal.

Although this workflow design fully leverages the generalized knowledge of the foundation models, eliminating dependence on scarce task-specific training data for vision-language navigation, several critical challenges persist. The approach relies on complex and fragile prompt engineering, it demands substantial computational resources for sequential prompting, and suffers from noise in generated captions and summaries – uncontrollable factors that collectively degrade information fidelity. Moreover, the capability of LLMs to accurately comprehend environmental spa-

tial structures and causal relationships for motion planning remain questionable, as evidenced by the nearly 40 % performance gap between LLM-based zero-shot navigation methods and specialized VLN models on R2R benchmarks. These limitations fundamentally constrain the reliability of text-exclusive foundation models for navigation tasks.

Complementary approaches have explored fine-tuning foundation models for vision-language navigation by encoding visual observations into either feature embeddings or textual inputs while reformulating navigation actions as structured text sequences to align with autoregressive prediction objectives. Although these methods aim to harness the pretrained capabilities of large models, and their performance remains substantially inferior to specialized VLN systems – a gap potentially attributable to either insufficient training data for activating the models' full potential or fundamental misalignment between the pretraining objectives and core VLN requirements, such as cross-view observation-instruction grounding in dynamic environments. More critically, such fine-tuning may degrade the models' inherent general-purpose linguistic competence, exacerbating key challenges in embodied AI by replacing interpretable, interactive language understanding with opaque black box behaviors that lack reliable spatial reasoning capabilities.

To address these challenges, we establish a novel manifold-aware historical topological mapping framework for navigation, grounded in differential geometry and deep learning theory. Building upon InstructBLIP's architecture, our method enhances visual-language adaptation through multi-image perception and chain-of-thought reasoning, which generates reliable executable navigation instructions while minimizing hallucination-induced failures. The proposed multi-curvature manifold feature mapping analyzes object distribution impacts by continuously deforming averaged visual feature vectors across curvature-variant spaces, leveraging their inherent geometric properties to capture hierarchical topological structures of diverse shapes. By jointly decoding the vision-language model's latent representations with topological graph embeddings, our approach enables long-term navigation history tracking and efficient backtracking while preserving the foundation model's core linguistic capabilities. The contribution of this paper can be summarized as:

- We address the long-horizon reasoning challenge by augmenting the encoder-decoder foundation model with historical observation encoding to provide global topological guidance.
- We mitigate hallucination in model reasoning through multi-curvature manifold learning that identifies spatiotemporal variations in visual sequences, enabling reference-based situational awareness and decision-making.
- We resolve the misalignment between instruction segments and dynamic environments via a graph self-attention mechanism that jointly models node proximity and visual similarity in historical topological maps to establish spatially-grounded relationships.

2 RELATED WORKS

2.1 Vision-and-Language Navigation (VLN)

Visual-language navigation (VLN) tasks demand intelligent agents to either reach specified locations [6, 7] or locate target objects [8, 9] in photorealistic environments following natural language instructions, whose practical applications have stimulated considerable research interest as they represent fundamental competencies for embodied AI systems, with early approaches employing recurrent neural network architectures [10, 11, 12, 13] before transformer-based models [14, 15, 16, 17, 18] achieved notable advancements through superior long-range dependency modeling, though the limited scale of VLN datasets introduces substantial bias and overfitting risks that subsequent methods attempted to mitigate through speaker–follower paradigms generating synthetic instructions [19, 20, 21, 22, 23], cross-dataset trajectory-instruction pair collection [24, 25, 26], or environment editing techniques [27, 28], while such solutions still cannot fully eliminate dataset-induced biases, prompting our proposed causal inference framework that explicitly models cause–effect relationships during training to enhance adaptation to distributional variations.

2.2 Large Language Models in VLN

The integration of Large Language Models (LLMs) into robotic navigation systems has gained significant attention due to their advanced linguistic comprehension, inherent communicative capabilities, and robust commonsense reasoning, with existing visual-language navigation (VLN) research primarily adopting three distinct LLM incorporation strategies: model ensemble approaches exemplified by MIC [29] employing GPT-2 [30] for auxiliary navigation guidance, zero-shot agent implementations like NavGPT [1] demonstrating basic navigational reasoning through sophisticated prompting systems alongside multi-agent frameworks (DiscussNav [2]) and topology-enhanced variants (MapGPT [4]), though these exhibit substantial performance deficits compared to supervised methods even when utilizing state-of-the-art models like GPT-4 [31], and specialized fine-tuning methodologies including LangNav [3] and NavCoT [32] adapting LLaMA-7B [33] to investigate linguistic perceptual representations, while NaviLLM [5] pursues multi-task optimization to develop generalist VLN agents.

2.3 Modal Alignment for Large Visual Language Models (VLMs)

Recent advances in Large Language Models (LLMs) have spurred significant interest in adapting these pretrained systems for multimodal applications, particularly in visual information processing and interpretation [34, 35, 36, 37, 38, 39, 40, 41, 42, 43], where two predominant methodologies have emerged for modality integration: the query-based visual resampling approach pioneered by Flamingo [34] and BLIP-2 [36],

later refined in MiniGPT-4 [39], InstructBLIP [37] and Qwen-VL [43], and the alternative strategy employing dense projection layers to directly align vision encoder outputs with LLM input spaces as demonstrated in LLaVA [35] and MiniGPT-v2 [40], with our implementation selecting the Q-former architecture from InstructBLIP [37] to optimally manage multi-view visual sequence length constraints during navigation tasks.

3 METHOD

3.1 Overall Framework Review

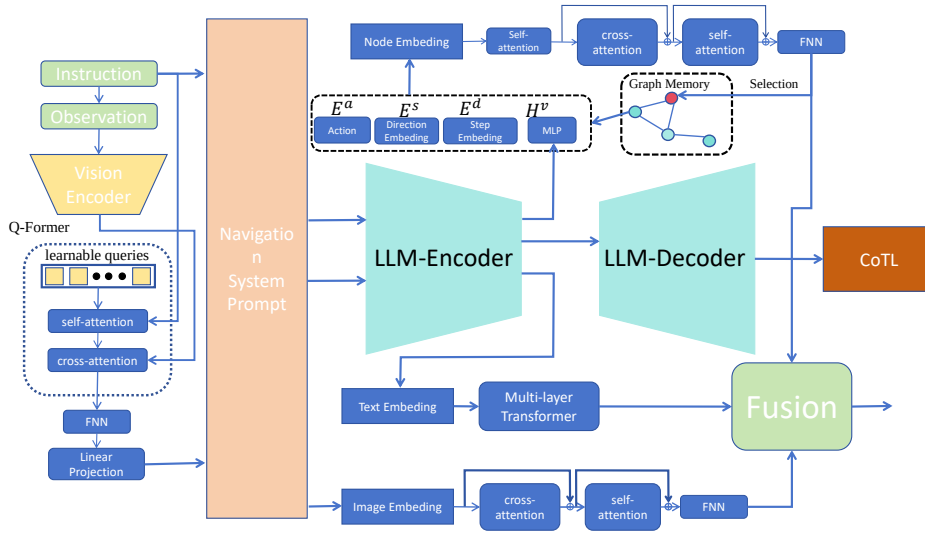


Figure 1. The proposed manifold-aware visual-language navigation framework. It processes visual observations and instructions through aligned tokenization, integrates historical topological context into a global graph, and fuses multi-modal features via residual attention to generate decisions while producing visually-grounded chain-of-thought reasoning.

The proposed framework integrates three core components – a VLM, an encoder-decoder foundation model, and a navigation policy network – as illustrated in Figure 1 with the vision-language module, visual observations and linguistic instructions are processed by the Q-former to generate compact image tokens that preserve spatial details while serving as visual inputs for the language model's reasoning pipeline. For navigation action prediction, the system utilizes latent representations derived from both the encoded image tokens and instruction text embeddings, which are jointly processed by the foundation model's encoder.

For a given instruction $\mathbb{W}$ composed of word embeddings that requires the agent to follow it for navigation in a predefined undirected graph $\mathbb{G} = \langle \mathbb{V}, \mathbb{E} \rangle$, where $\mathbb{V}$ represents navigable nodes and $\mathbb{E}$ denotes connectivity edges, at step t, the agent perceives its surroundings by observing a set of multi-view perspectives $\mathbb{O} = \langle o_i, angle_i \rangle$, $i = 1 \ldots N$ for each connected navigable node candidate, with N indicating the number of candidate nodes, each unique view represented as o_i, and its associated angular orientation relative to the agent's heading denoted as $angle_i$. The agent is required to predict subsequent actions by selecting an angle $angle_t$ relative to the visual observation o_t, and it must learn a navigation policy $\pi(angle_t|\mathbb{W}, O_t; \theta)$ under network parameters θ.

a) Object detection results b) Object segmentation results c) Integration of depth information

Figure 2. Visual feature processing pipeline for observation sequences

To efficiently encode multi-view images for spatial-aware navigation reasoning in large language models, a dual-branch network architecture is employed. The first branch utilizes the Q-former from InstructBLIP [37] to encode each view into fixed-length visual tokens, where learnable queries $\mathbb{Q}_i \in R^{32\times768}$ first perform self-attention with instruction embeddings ι before cross-attending to visual features extracted by the vision encoder, with the resulting queries linearly projected as image tokens for the LLM. The second branch enhances distance perception by matching object detection and target segmentation results through IoU computation, where pixels of the i^{th} object in RGB images are obtained via intersecting bounding boxes b_i and masks k_i before being mapped to depth maps to calculate average pixel values as object-agent distances, as shown in Figure 2. The proposed framework processes visual observations through the foundation model's encoder, preserving spatial details as feature vectors, while directional cues for candidate views are structurally embedded in navigation prompts using the format "Candidate i, facing $angle_i$, direction" to orient the language model, with special tokens <IMG></IMG> and <INST></INST> demarcating visual tokens and instructions respectively, followed by Q-former and projection layer fine-tuning on autoregressively generated navigation reasoning data from R2R training sets, ultimately yielding a complete vision-language model capable of generating detailed navigation reasoning through the LLM decoder.

3.2 Manifold-Based Visual Reference Reasoning with Multi-Curvature Spatiotemporal Awareness

Unlike conventional image or video captioning, actionable navigation instructions require not only visual descriptions but also references to environmental landmarks for directional guidance at critical turning points, with cognitive psychology studies revealing humans typically identify key navigational waypoints in mental maps before verbalizing paths, necessitating visual reference identification as a core capability for instruction generation; this work therefore introduces multi-curvature manifold-based spatiotemporal analysis to guide chain-of-thought reasoning in selecting trajectory-critical visual landmarks for instruction synthesis.

To address the potential incompatibility of cross-modal feature embeddings, we propose a comprehensive hybrid manifold space to enhance visual observation encoding, as shown in Figure 3, where curvature-specific manifolds – zero-curvature Euclidean space for universal patterns, positive-curvature spherical manifolds, and negative-curvature hyperbolic manifolds – are integrated to leverage complementary geodesic distance metrics for more expressive feature relationships. A hyperbolic manifold is a Riemannian manifold with constant negative Gaussian curvature, where the exponentially expanding space encodes more information than Euclidean space at equal dimensionality. This property stems from the quadratic growth of reachable network space per radial step from center nodes, mirroring hierarchical VLN structures, e.g., building-floor-room-object. Hyperbolic embeddings enable agents to infer path-forking logic through geometric relationships, while their distance sensitivity and gradient properties alleviate long-tail distribution challenges in instruction-path data. The manifold's metric adjustments enhance low-frequency sample representation, and its geometric alignment capability jointly embeds tree-like language instructions with hierarchical visual scenes via shared curvature parameters. A positively curved manifold exhibits positive curvature in all directions at every point, exemplified by a sphere where space grows more slowly than Euclidean space with increasing radius due to its compact geometry. Such manifolds naturally handle directional data like camera poses or lighting directions, making them ideal for VLN's orientation-specific commands by preserving angular continuity,for example, eliminating Euclidean discontinuity between 0° and 360°. Their inherent compactness also aligns with finite navigation spaces, justifying multi-curvature space integration for environmental observation. The fusion of multi-curvature spaces enables the model to adaptively select the most appropriate geometric representations, thereby capturing the heterogeneous spatial and semantic relationships in complex navigation environments more comprehensively and robustly. This further justifies the integration of multi-curvature spaces for environmental observation.

For instruction–path pairs in the training set, we first extract nouns from instructions to form the initial visual reference set Λ_x. Since annotated instructions may omit critical references, we augment Λ_x by emulating human path observation patterns that select salient references through temporal and spatial perspectives. Temporally, humans remember visually distinctive scene transitions (e.g.,

Figure 3. Visual reference selection (left) and multi-curvature manifold reasoning (right): Candidate views display objects grouped in blue boxes, while only action-critical objects are selected as visual references (red boxes). Temporal selection identifies viewpoint transitions causing significant scene changes (e.g., all three illustrated actions alter the global context).

moving from living room to bedroom) – we thus compute feature differences between panoramic views along trajectories to identify such transition points. Given a path, we construct a sequence of mean-pooled panoramic features $\{V_t^*\}_{t=1}^T$ and measure temporal importance scores δ_t^τ via cosine distance between V_t^* and V_{t+1}^*:

$$\delta^t = 1 - \frac{V_t^* \cdot V_{t+1}^*}{\|V_t^*\| \cdot \left\|V_{t+1}^*\right\|}, V_t^* = \frac{1}{K}\sum_{k=1}^{K} V_{t,k}, \tag{1}$$

where K denotes the view index and t represents the timestep. From a spatial perspective, we identify distinctive objects as visual references by selecting those appearing exclusively in the chosen action view V_{t,a_t} at timestep t but not in other candidate views. We first extract all objects from V_{t,a_t} as candidate references, then assign uniqueness scores based on their presence in alternative candidate views o_1, o_2, o_3. For instance, visual reference $\lambda_{t,n}^*$ appearing in these views receives spatial importance score $\delta_{t,n}^a$:

$$\delta_{t,n}^a = 1 - d_{t,o_1}^a - d_{t,o_2}^a - d_{t,o_3}^a, \tag{2}$$

where d_{t,c_1}^{angle} represents the cosine similarity between the visual observation of the selected action view and candidate view o_i. For a candidate visual reference object, its final spatiotemporal distinctiveness score is calculated as:

$$\delta_{t,n} = \delta_{t,n}^a \cdot \delta_t^\tau. \tag{3}$$

We retain visual references with spatiotemporal distinctiveness scores $\delta_{t,n}$ exceeding a threshold to form a supplemental set Λ_v, yielding the complete landmark set $\Lambda = \Lambda_x \cup \Lambda_v$. Real-world environments exhibit feature space embedding inconsistencies due to structural complexity – while Euclidean space effectively represents

regular objects (chairs, tables), hyperbolic space better captures hierarchical structures (sofas, hourglasses), and spherical space suits fine-grained circular objects. To address this, we employ multi-curvature manifolds for optimal spatiotemporal representation. For zero-curvature Euclidean space, encoder Φ directly extracts feature vectors X from multi-view observations:

$$f_{euc} = \Phi(X). \tag{4}$$

Here, f_{euc} represents the features extracted in Euclidean space, which preserves the original data structure and is suitable for regular objects in the environment. Hyperbolic space, characterized by constant negative curvature, employs exponential mapping for feature projection. Specifically, given features extracted by encoder Φ, we compute their norm and apply a hyperbolic tangent function with appropriate scaling to achieve hyperbolic space mapping:

$$f_{hyp} = tanh(\sqrt{C}||\Phi(X)||) \cdot \frac{\Phi(X)}{\sqrt{C}||\Phi(X)||}, \tag{5}$$

where $-C$ denotes the curvature of hyperbolic space (typically set to -1) and f_{hyp} represents the extracted hyperbolic features, which are particularly effective for modeling hierarchical structures and non-linear relationships in visual-language navigation environments.

Spherical space, characterized by constant positive curvature, requires feature vectors to lie on the unit sphere's surface. This is achieved by normalizing features from encoder Φ, followed by spherical projection function $proj$:

$$f_{sph} = proj\left(\frac{\Phi(X)}{||\Phi(X)||}\right), \tag{6}$$

where f_{sph} denotes the features extracted in spherical space, with the spherical projection function $proj$ implemented as:

$$radius = \sqrt{\varpi_x^2 + \varpi_y^2 + \varpi_z^2}, \tag{7}$$

$$\partial_1 = \arccos\left(\frac{\varpi_z}{radius}\right), \partial_2 = \text{atan2}(\varpi_y, \varpi_x), \tag{8}$$

where ϖ_x, ϖ_y, and ϖ_z represent the coordinates along the spherical space's three axes, $radius$ denotes the distance from the origin in spherical mapping, ∂_1 indicates the latitude on the sphere, and ∂_2 specifies the longitude.

After obtaining visual mapping features from all three curvature spaces, we compute comprehensive spatiotemporal distinctiveness scores across multi-curvature manifolds following the visual reference selection framework:

$$\delta_{t,n,all} = \delta_{t,n}^{a} \cdot \delta_{t}^{\tau} + \delta_{t,n,hyp}^{a} \cdot \delta_{t,hyp}^{\tau} + \delta_{t,n,sph}^{a} \cdot \delta_{t,sph}^{\tau}. \tag{9}$$

3.3 Topology-Aware Navigation Policy Network with Historical Structure Encoding

Prior research reveals that fine-tuning large language models for vision-language navigation fundamentally suffers from their deficient spatial understanding and limited long-range reasoning capacity during navigation tasks, which we address through a topological graph-based navigation strategy illustrated in Figure 4 where the dynamically maintained graph serves as an experiential memory module, enabling the language model to select optimal forward actions from the complete topological structure while supporting efficient backtracking to unexplored nodes upon encountering erroneous paths.

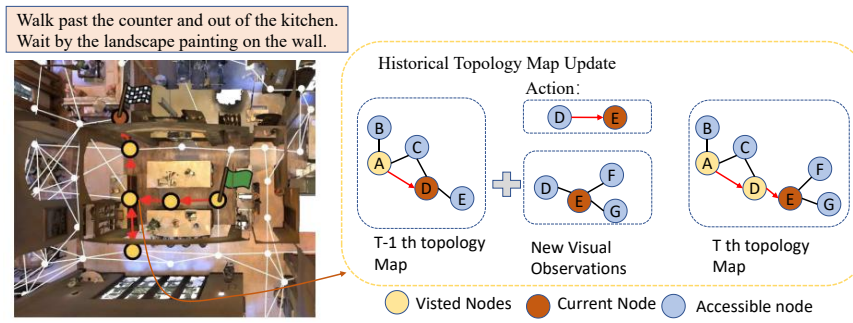


Figure 4. Historical trajectory graph of the topology. Given a new action representative moving from node D to node E, the agent receives a new visual observation at node E. The new node is designated as the current node and its representation is updated accordingly.

The environment graph $\mathbb{G}$ is initially unknown to the agent, so our method incrementally constructs $\mathbb{G}_t = \langle \mathbb{V}_t, \mathbb{E}_t \rangle$ with K_t nodes through navigational observations after t steps. $\mathbb{V}_t$ contains three node types: visited nodes (with full panoramic views), navigable nodes (partially observed from visited locations), and the current node. At each step t, the model updates the graph by adding the current node and adjacent unexplored nodes while refreshing edge connections and visual representations for both current and navigable nodes based on new observations.

The historical memory topology graph stores visited nodes and adjacent unexplored nodes along the trajectory. Our method represents each visited node using average-pooled features of all its candidate views, while unexplored nodes are represented by partial pooling of corresponding views from their neighboring visited nodes. Each view is characterized by its visual features $\bar{H}_i^{'v}$ and directional embedding E^d, which encodes the node's egocentric position relative to the current node in terms of orientation and distance. The step embedding E^s indicates the traversal order, marking visited nodes with their latest timestep and unexplored nodes with 0. The action token E^a flags whether the view was previously selected as a movement target. A dedicated "stop" node connects to all others to represent

termination. This encoding scheme improves instruction alignment by capturing diverse navigation histories. A multi-layer Transformer models spatial relationships between nodes:

$$\mathbb{V} = selfAttn\left(\frac{1}{M}\sum_{i=1}^{M}\left(\bar{H_i'}^{v} + E_i^d + E_i^s + E_i^a\right)\right), \tag{10}$$

where M denotes the number of views per node. The navigation graph at step t, represented as $\mathbb{G}_t = \langle \mathbb{V}_t, \mathbb{E}_t \rangle$, processes its node embeddings $\mathbb{V}_t$ through a multi-layer cross-modal Transformer that models instruction-node relationships by first performing cross-attention with the LLM-encoded instructions, followed by Graph-Aware Self-Attention (GASA) that fuses semantic information through joint consideration of inter-node distances and visual similarities:

$$GASA(\mathbb{V}) = softmax\left(\frac{\mathbb{V}W_q(\mathbb{V}W_k)^T}{\sqrt{d}} + (\mathbb{E}w_e + b_e)\right)\mathbb{V}W_v. \tag{11}$$

The proposed framework employs a two-layer feedforward network to process the GASA-generated node representations and produce action scores, where $\mathbb{V}$ represents the encoded node features, $\mathbb{E}$ denotes the pairwise distance matrix extracted from $\mathbb{G}_t = \mathbb{V}_t, \mathbb{E}_t$, and w_e and b_e correspond to learnable parameters. The agent navigates by selecting the highest-scoring unvisited node as the target while following the shortest path in the graph memory, with score masking applied to previously visited nodes to promote exploration. A separate two-layer feedforward network performs multi-branch feature fusion.

Methods	Val Unseen				
	TL	NE	OSR	SR	SPL
NavGPT (GPT-4, AAAI2024) [1]	11.45	6.46	42	34	29
MapGPT (GPT-4, ACL2024) [4]	–	6.92	58	39	26
DiscussNav (GPT-4, ICRA2024) [2]	9.69	5.32	61	43	40
NavCoT (LLAMA2-7B, TPAMI2025) [32]	9.95	6.26	48	40	37
LangNav (LLAMA2-7B, ACL2024) [3]	–	–	–	46	–
NaviLLM (Vicuna-7B, CVPR2024) [5]	12.81	3.51	–	67	59
Our Method (1.5B)	13.26	3.03	80	73	60
Our Method (5B)	12.37	**2.92**	**82**	**74**	**63**

Table 1. Comparison with large model-based navigation methods on the R2R dataset

3.4 Multi-Stage Learning for Action Inference via Imitation Learning and Data Aggregation

The proposed method employs a two-stage training approach for learning the LLM's action prediction and navigation reasoning generation. In the first stage, the model is initialized from InstructBLIP [37] and follows standard multimodal foundation

Methods	Val Seen					Val Unseen				
	TL	NE	OSR	SR	SPL	TL	NE	OSR	SR	SPL
Navigation Methods Using Vision-Language-Action Pre-Training Tasks										
PREVALENT (CVPR2020) [14]	10.32	3.67	–	69	65	10.19	4.71	–	58	53
AirBERT (ICCV2021) [25]	11.09	2.68	–	75	70	11.78	4.10	–	62	56
VLN BERT (CVPR2021) [15]	11.13	2.90	–	72	68	12.01	3.93	–	63	57
MARVAL (CVPR2023) [23]	10.60	2.99	–	73	69	10.15	4.06	–	65	61
HAMT (NIPS2021) [16]	11.15	2.51	82	76	72	11.46	2.29	73	66	61
DUET (CVPR2022) [17]	–	2.28	86	79	73	13.94	3.31	81	72	60
HOP+ (TPAMI2023) [44]	11.31	2.33	–	78	73	11.76	3.49	–	67	61
BEVBert (ICCV2023) [45]	13.56	2.17	88	81	74	14.55	2.81	84	75	64
DUET + ScaleVLN (ICCV2023) [46]	11.90	2.16	87	80	75	12.40	2.34	87	**79**	**70**
GOAT (CVPR2024) [47]	–	1.79	88	**83**	**79**	–	2.40	84	77	68
Navigation Methods Without Vision-Language-Action Pre-Training Tasks										
DUET (w/o pretrain) [17]	12.38	3.62	73	66	60	13.20	4.07	72	64	55
Our Method (1.5B)	12.63	2.73	80	74	66	13.26	3.03	80	73	60
Our Method (5B)	12.58	2.19	84	79	70	12.37	2.92	82	74	63

Table 2. Comparison with large-scale pre-training based methods on the R2R dataset

model fine-tuning procedures, where only the Q-former is fine-tuned on collected navigation reasoning data while keeping the LLM and visual encoder frozen. The second stage connects the pretrained VLM with the downstream navigation policy network, fine-tuning solely the policy network with the VLM remaining frozen. For policy network fine-tuning, we adopt established practices by combining imitation learning and data aggregation losses, defining the imitation learning loss function $\mathbb{L}_{bc}$ as:

$$\mathbb{L}_{bc} = -\sum_{t=1}^{T} \log \pi(a_t^* | \iota, \mathcal{G}_t), \tag{12}$$

where a_t^* denotes the ground-truth decision. The data aggregation loss function $\mathbb{L}_{DAG}$ is defined as:

$$\mathbb{L}_{DAG} = -\sum_{t=1}^{T} \log \pi(\bar{a_t^*}|\iota, \bar{\mathcal{G}_t^*}), \tag{13}$$

where ι represents the encoded instruction text features and $\bar{a_t^*}$ denotes the pseudo-action labels determined by the shortest path from the partially constructed graph $\bar{\mathcal{G}_t^*}$ (generated through policy-based upsampling) to the destination.

$$\mathbb{L} = \varsigma \mathbb{L}_{bc} + \mathbb{L}_{DAG}. \tag{14}$$

The total training loss $\mathbb{L}$ combines both components with ς as the balancing factor.

4 EXPERIMENT

4.1 Dataset Introduction and Processing

The Room-to-Room (R2R) dataset, collected from the Matterport3D simulator [6], comprises 10 800 panoramic views, 21 567 navigation instructions, and 7 189 corresponding trajectories across 90 realistic building-scale indoor environments, with each trajectory annotated by three distinct instructions. The dataset is partitioned into four subsets: training set (train), validation set in seen environments (validation seen), validation set in unseen environments (validation unseen), and test set (test).

Metric: Trajectory Length (TL) measures the average path length in meters. Navigation Error (NE) calculates the mean distance between final and target positions. Success Rate (SR) represents the percentage of paths with Navigation Error below 3 meters. Oracle Success Rate (OSR) indicates the success rate under ideal stopping conditions. Success Rate penalized by Path Length (SPL) evaluates both success and efficiency by incorporating path length into the metric.

4.2 Experimental Setup

We employ InstructBLIP [37] as the foundation model component, evaluating both FlanT5-XL (3B) and FlanT5-XXL (11B) variants with a ViT-g/14 visual encoder. Phase one training initializes from pretrained InstructBLIP weights, processing 200 000 steps with batch size 8 using AdamW optimization ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.05), featuring linear learning rate warmup from 10^{-8} to 10^{-5} over 1 000 steps followed by cosine decay. Phase two fine-tunes the policy network with frozen vision-language model (batch size 2, lr = 10^{-5}). Implementation uses PyTorch 2.2.0 + cu117 on Ubuntu 20.04.2 LTS with NVIDIA RTX 4090 GPUs, Intel Xeon Platinum 8350C CPUs, and 128 GB RAM.

Methods	Val Seen					Val Unseen				
	TL	NE	OSR	SR	SPL	TL	NE	OSR	SR	SPL
Using the original 10 % R2R training data										
DUET [17]	12.75	6.18	54	44	38	13.08	5.78	58	48	40
Our Method (1.5B)	20.62	5.46	70	50	35	20.93	5.13	75	52	36
Our Method (5B)	20.12	3.85	79	64	46	19.31	3.93	80	65	47
Using the original 50 % R2R training data										
DUET [17]	13.46	4.65	68	56	49	13.57	4.41	70	60	50
Our Method (1.5B)	17.96	3.44	80	68	55	16.8	3.35	82	69	53
Our Method (5B)	15.38	2.85	81	73	61	15.36	2.86	84	75	61
Using the original 100 % R2R training data										
DUET [17]	12.38	3.62	73	66	60	13.20	4.07	72	64	54
Our Method (1.5B)	12.34	3.44	75	70	62	12.32	3.64	75	68	58
Our Method (5B)	12.48	2.69	81	74	67	13.17	3.15	80	72	61

Table 3. Ablation experiments on the R2R dataset on the effect of training data volume on model performance

4.3 Comparative Experiment

As shown in Table 1, our method compares with all existing foundation model-based VLN approaches for indoor environments, where we evaluate two InstructBLIP variants: FlanT5-XL (1.5B) and FlanT5-XXL (5B) encoder-only configurations (half the full model size since navigation decisions require only the encoder, though the complete model is needed for reasoning chain generation). Notably, our 5B model surpasses NaviLLM [5] by 7 % in success rate despite comparable or smaller parameter counts relative to competitors. Like other frozen-LLM approaches – including explicit-reasoning NavGPT [1], map-guided MapGPT [4], and multi-agent DiscussNav [2], our method preserves the LLM's native linguistic capabilities while maintaining model-swappability, interpretable path reasoning, and superior unseen-scene generalization.

As shown in Table 2, our method gets results that are close to those of methods with pretraining. These other methods use extra proxy tasks. The top two methods – ScaleVLN [46] and GOAT [47] – need a lot of extra data. ScaleVLN uses 1 000 large HM3D [48] and Gibson [49] scenes. GOAT uses ENVedit [28] data with causal learning. On the other hand, our training uses only the base R2R dataset and PREVALENT [14] pretrained features. These features are the same starting point for all methods before they add more data.

Notably, when pretrained VLN models abandon their proxy tasks and train directly on navigation tasks (e.g., DUET [17], an influential work in the field), performance drops significantly – 13 % in seen and 8 % in unseen scenarios. Our approach deliberately avoids additional pretraining tasks, instead focusing on fully activating the foundation model's inherent capabilities without introducing extra data, tasks, or computational costs.

Methods	Val Seen					Val Unseen				
	TL	NE	OSR	SR	SPL	TL	NE	OSR	SR	SPL
Our Method (1.5B)	12.63	2.73	80	74	66	13.26	3.03	80	73	60
w/o history	33.99	8.97	70	24	9	34.55	9.47	72	18	6
w/o local image	12.94	2.94	80	73	64	13.78	3.14	80	71	59
w/o Q-former	15.83	3.16	83	72	60	15.69	3.32	82	70	55
w/o manifold	15.37	3.15	83	70	59	15.4	3.40	80	69	55

Table 4. Ablation experiments on the R2R dataset on the impact of individual modules on model performance

4.4 Ablation Experiment

Data scarcity remains a critical bottleneck for visual-language navigation performance. To demonstrate our method's data efficiency, we conduct ablation studies on R2R's original dataset with varying training data proportions, revealing the proposed approach's few-shot learning capability when activating foundation model potentials (Table 3). The performance gap between our method and DUET [17] widens with reduced data, while differing by only 8 % (success rate) at full data utilization with our 5B model, the gap expands to 20 % when using merely 10 % training data. This progression highlights the advantage of foundation model-based approaches in a few/zero-shot learning scenarios. The ablation studies in Table 4 validate each module's contribution by systematically removing components. Removing the historical topology encoding and vision-language cross-attention, labeled as w/o history and replaced with an MLP in Figure 1's upper section, causes a 50 % success rate drop, proving that relying solely on foundation models for environmental understanding and directional decisions is ineffective. Omitting raw image features, marked as w/o local image from Figure 1's lower branch, has a marginal impact on performance despite providing valuable visual details. Disabling the first-stage Q-former training, noted as w/o Q-former, slightly reduces success rates due to its critical role in aligning spatial descriptions with visual landmarks. The most significant performance decline occurs when eliminating the multi-curvature manifold-based reference filtering, labeled as w/o manifold, confirming the necessity of explicitly identifying discriminative visual-linguistic references for observation changes.

4.5 Visualization of the Reasoning Process

The visualization in Figures 5 and 6 presents the large language model's reasoning process aligned with panoramic visual observations, explicitly showing the model's perceived visual content, its interpretation of the currently executed instruction segments, and its future planning. By specially coloring the visual references that the model focuses on during reasoning, we demonstrate its reference-centered sentence structure that standardizes the model's outputs. The figures reveal that the model first analyzes and describes the current environment by identifying key visual ref-

Figure 5. Large model reasoning visualization process Step 0 – Step 2

Figure 6. Large model reasoning visualization process Step 3 – Step 6

erences in the front, right, and left views, anchoring its descriptions around these references to minimize hallucination effects while maintaining alignment with ongoing instruction execution.

The navigation instruction shown in the visualization divides into three major environments and three main steps: starting from outside the bedroom and turning right until facing the bedroom and a bed, then walking toward the bed to enter the bedroom before turning right toward the bedroom door, followed by turning right to enter the hall and then turning left to reach the stairway. The environmental transitions progress from outside the bedroom to inside the bedroom and finally to the hall. The large model exhibits clear phase awareness during reasoning, accurately recognizing which steps have been completed and what actions need to be taken next.

5 CONCLUSION

This paper proposes and validates a manifold-aware navigation method for large models guided by historical topological graphs. First, we leverage the image-text representation capability of large models using an encoder-decoder architecture, where the encoder is enhanced with historical topological structure encoding to support long-range reasoning. To address hallucination effects in stepwise reasoning, where the model may generate descriptions of non-existent content, we introduce a visual reference screening method based on multi-curvature spatiotemporal manifold divergence. This ensures the model remains focused on relevant reasoning details. A case study visually demonstrates how our approach guides the model to perform navigation reasoning anchored on visual references. Experimental results validate the effectiveness of the proposed method and quantify the contributions of its individual components. In future work, we plan to extend the framework to outdoor navigation scenarios and investigate the integration of real-time sensor inputs (e.g., LiDAR, IMU) for more robust embodied reasoning.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62571464, the Shenzhen Science and Technology Projects under Grant No. JCYJ20200109143035495, and the Natural Science Foundation of Fujian Province under Grants No. 2022J011275 and No. 2023J01003.

REFERENCES

[1] Zhou, G.—Hong, Y.—Wu, Q.: NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, No. 7, pp. 7641–7649, doi: 10.1609/aaai.v38i7.28597.

[2] Long, Y.—Li, X.—Cai, W.—Dong, H.: Discuss Before Moving: Visual Language Navigation via Multi-Expert Discussions. 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 17380–17387, doi: 10.1109/ICRA57147.2024.10611565.

[3] Pan, B.—Panda, R.—Jin, S.—Feris, R.—Oliva, A.—Isola, P.—Kim, Y.: LangNav: Language as a Perceptual Representation for Navigation. In: Duh, K., Gomez, H., Bethard, S. (Eds.): Findings of the Association for Computational Linguistics: NAACL 2024. 2024, pp. 950–974, doi: 10.18653/v1/2024.findings-naacl.60.

[4] Chen, J.—Lin, B.—Xu, R.—Chai, Z.—Liang, X.—Wong, K. Y. K.: MapGPT: Map-Guided Prompting for Unified Vision-and-Language Navigation. In: Ku, L. W., Martins, A., Srikumar, V. (Eds.): Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2024). 2024, pp. 9796–9810, doi: 10.18653/v1/2024.acl-long.529.

[5] Zheng, D.—Huang, S.—Zhao, L.—Zhong, Y.—Wang, L.: Towards Learning a Generalist Model for Embodied Navigation. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13624–13634, doi: 10.1109/CVPR52733.2024.01293.

[6] Anderson, P.—Wu, Q.—Teney, D.—Bruce, J.—Johnson, M.—Sünderhauf, N.—Reid, I.—Gould, S.—van den Hengel, A.: Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3674–3683, doi: 10.1109/CVPR.2018.00387.

[7] Ku, A.—Anderson, P.—Patel, R.—Ie, E.—Baldridge, J.: Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.): Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020). ACL, 2020, pp. 4392–4412, doi: 10.18653/v1/2020.emnlp-main.356.

[8] Qi, Y.—Wu, Q.—Anderson, P.—Wang, X.—Wang, W. Y.—Shen, C.—van den Hengel, A.: REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9979–9988, doi: 10.1109/CVPR42600.2020.01000.

[9] Zhu, F.—Liang, X.—Zhu, Y.—Yu, Q.—Chang, X.—Liang, X.: SOON: Scenario Oriented Object Navigation with Graph-Based Exploration. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12689–12699, doi: 10.1109/CVPR46437.2021.01250.

[10] Wang, X.—Huang, Q.—Celikyilmaz, A.—Gao, J.—Shen, D.—Wang, Y. F.—Wang, W. Y.—Zhang, L.: Vision-Language Navigation Policy Learning and Adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, 2021, No. 12, pp. 4205–4216, doi: 10.1109/TPAMI.2020.2972281.

[11] An, D.—Qi, Y.—Huang, Y.—Wu, Q.—Wang, L.—Tan, T.: Neighbor-View Enhanced Model for Vision and Language Navigation. Proceedings of the 29th ACM International Conference on Multimedia (MM '21), 2021, pp. 5101–5109, doi: 10.1145/3474085.3475282.

[12] Dang, R.—Shi, Z.—Wang, L.—He, Z.—Liu, C.—Chen, Q.: Unbiased Directed Object Attention Graph for Object Navigation. Proceedings of the 30th

ACM International Conference on Multimedia (MM '22), 2022, pp. 3617–3627, doi: 10.1145/3503161.3547852.

[13] He, Z.—Wang, L.—Li, S.—Yan, Q.—Liu, C.—Chen, Q.: A Multi-Level Attention Network with Sub-Instruction for Continuous Vision-and-Language Navigation. Applied Intelligence, Vol. 55, 2025, No. 7, Art. No. 657, doi: 10.1007/s10489-025-06544-9.

[14] Hao, W.—Li, C.—Li, X.—Carin, L.—Gao, J.: Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-Training. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13134–13143, doi: 10.1109/CVPR42600.2020.01315.

[15] Hong, Y.—Wu, Q.—Qi, Y.—Rodriguez-Opazo, C.—Gould, S.: VLN-BERT: A Recurrent Vision-and-Language BERT for Navigation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1643–1653, doi: 10.1109/CVPR46437.2021.00169.

[16] Chen, S.—Guhur, P. L.—Schmid, C.—Laptev, I.: History Aware Multimodal Transformer for Vision-and-Language Navigation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., Wortman Vaughan, J. (Eds.): Advances in Neural Information Processing Systems 34 (NeurIPS 2021). Curran Associates, Inc., 2021, pp. 5834–5847, https://proceedings.neurips.cc/paper_files/paper/2021/file/2e5c2cb8d13e8fba78d95211440ba326-Paper.pdf.

[17] Chen, S.—Guhur, P. L.—Tapaswi, M.—Schmid, C.—Laptev, I.: Think Global, Act Local: Dual-Scale Graph Transformer for Vision-and-Language Navigation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16516–16526, doi: 10.1109/CVPR52688.2022.01604.

[18] Wang, L.—He, Z.—Tang, J.—Dang, R.—Wang, N.—Liu, C.—Chen, Q.: A Dual Semantic-Aware Recurrent Global-Adaptive Network for Vision-and-Language Navigation. In: Elkind, E. (Ed.): Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23), Main Track. 2023, pp. 1479–1487, doi: 10.24963/ijcai.2023/164.

[19] Fried, D.—Hu, R.—Cirik, V.—Rohrbach, A.—Andreas, J.—Morency, L. P.—Berg-Kirkpatrick, T.—Saenko, K.—Klein, D.—Darrell, T.: Speaker-Follower Models for Vision-and-Language Navigation. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 31 (NeurIPS 2018). Curran Associates, Inc., 2018, pp. 3314–3325, https://proceedings.neurips.cc/paper_files/paper/2018/file/6a81681a7af700c6385d36577ebec359-Paper.pdf.

[20] Tan, H.—Yu, L.—Bansal, M.: Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL 2019), 2019, pp. 2610–2621, doi: 10.18653/v1/N19-1268.

[21] Wang, L.—Liu, C.—He, Z.—Li, S.—Yan, Q.—Chen, H.—Chen, Q.: PASTS: Progress-Aware Spatio-Temporal Transformer Speaker for Vision-and-Language Navigation. Engineering Applications of Artificial Intelligence, Vol. 128, 2023,

Art. No. 107487, doi: 10.1016/j.engappai.2023.107487.

[22] Wang, L.—He, Z.—Dang, R.—Chen, H.—Liu, C.—Chen, Q.: RES-StS: Referring Expression Speaker via Self-Training with Scorer for Goal-Oriented Vision-Language Navigation. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 33, 2023, No. 7, pp. 3441–3454, doi: 10.1109/TCSVT.2022.3233554.

[23] Kamath, A.—Anderson, P.—Wang, S.—Koh, J. Y.—Ku, A.—Waters, A.—Yang, Y.—Baldridge, J.—Parekh, Z.: A New Path: Scaling Vision-and-Language Navigation with Synthetic Instructions and Imitation Learning. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10813–10823, doi: 10.1109/CVPR52729.2023.01041.

[24] Majumdar, A.—Shrivastava, A.—Lee, S.—Anderson, P.—Parikh, D.—Batra, D.: Improving Vision-and-Language Navigation with Image-Text Pairs from the Web. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (Eds.): Computer Vision – ECCV 2020. Springer, Cham, Lecture Notes in Computer Science, Vol. 12351, 2020, pp. 259–274, doi: 10.1007/978-3-030-58539-6_16.

[25] Guhur, P. L.—Tapaswi, M.—Chen, S.—Laptev, I.—Schmid, C.: Airbert: In-Domain Pretraining for Vision-and-Language Navigation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1614–1623, doi: 10.1109/ICCV48922.2021.00166.

[26] Lin, K.—Chen, P.—Huang, D.—Li, T. H.—Tan, M.—Gan, C.: Learning Vision-and-Language Navigation from YouTube Videos. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8283–8292, doi: 10.1109/ICCV51070.2023.00764.

[27] Liu, C.—Zhu, F.—Chang, X.—Liang, X.—Ge, Z.—Shen, Y. D.: Vision-Language Navigation with Random Environmental Mixup. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1624–1634, doi: 10.1109/ICCV48922.2021.00167.

[28] Li, J.—Tan, H.—Bansal, M.: Envedit: Environment Editing for Vision-and-Language Navigation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15386–15396, doi: 10.1109/CVPR52688.2022.01497.

[29] Qiao, Y.—Qi, Y.—Yu, Z.—Liu, J.—Wu, Q.: March in Chat: Interactive Prompting for Remote Embodied Referring Expression. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15712–15721, doi: 10.1109/ICCV51070.2023.01444.

[30] Radford, A.—Wu, J.—Child, R.—Luan, D.—Amodei, D.—Sutskever, I.: Language Models Are Unsupervised Multitask Learners. OpenAI Blog, Vol. 1, 2019, No. 8, Art. No. 9, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[31] OpenAI—Achiam, J.—Adler, S.—Agarwal, S. et al.: GPT-4 Technical Report. CoRR, 2023, doi: 10.48550/arXiv.2303.08774.

[32] Lin, B.—Nie, Y.—Wei, Z.—Chen, J.—Ma, S.—Han, J.—Xu, H.—Chang, X.—Liang, X.: NavCoT: Boosting LLM-Based Vision-and-Language Navigation via Learning Disentangled Reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 47, 2024, No. 7, pp. 5945–5957, doi:

10.1109/TPAMI.2025.3554559.

[33] Touvron, H.—Lavril, T.—Izacard, G.—Martinet, X.—Lachaux, M. A.—Lacroix, T.—Rozière, B.—Goyal, N.—Hambro, E.—Azhar, F.—Rodriguez, A.—Joulin, A.—Grave, E.—Lample, G.: LLaMA: Open and Efficient Foundation Language Models. CoRR, 2023, doi: 10.48550/arXiv.2302.13971.

[34] Alayrac, J. B.—Donahue, J.—Luc, P.—Miech, A.—Barr, I. et al.: Flamingo: A Visual Language Model for Few-Shot Learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.): Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Curran Associates, Inc., 2022, pp. 23716–23736, https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

[35] Liu, H.—Li, C.—Wu, Q.—Lee, Y. J.: Visual Instruction Tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (Eds.): Advances in Neural Information Processing Systems 36 (NeurIPS 2023). Curran Associates, Inc., 2023, pp. 34892–34916, https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

[36] Li, J.—Li, D.—Savarese, S.—Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (Eds.): Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research (PMLR), Vol. 202, 2023, pp. 19730–19742, https://proceedings.mlr.press/v202/li23q.html.

[37] Dai, W.—Li, J.—Li, D.—Tiong, A.—Zhao, J.—Wang, W.—Li, B.—Fung, P. N.—Hoi, S. C.: InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (Eds.): Advances in Neural Information Processing Systems 36 (NeurIPS 2023). Curran Associates, Inc., 2023, pp. 49250–49267, https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf.

[38] Driess, D.—Xia, F.—Sajjadi, M. S. M.—Lynch, C.—Chowdhery, A. et al.: PaLM-E: An Embodied Multimodal Language Model. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (Eds.): Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research (PMLR), Vol. 202, 2023, pp. 8469–8488, https://proceedings.mlr.press/v202/driess23a.html.

[39] Zhu, D.—Chen, J.—Shen, X.—Li, X.—Elhoseiny, M.: MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (Eds.): International Conference on Learning Representations 2024 (ICLR 2024). 2024, pp. 18378–18394, https://proceedings.iclr.cc/paper_files/paper/2024/file/50623630a2372839c078474efa6c0cb8-Paper-Conference.pdf.

[40] Chen, J.—Zhu, D.—Shen, X.—Li, X.—Liu, Z.—Zhang, P.—Krishnamoorthi, R.—Chandra, V.—Xiong, Y.—Elhoseiny, M.: MiniGPT-V2: Large Language Model as a Unified Interface for Vision-Language Multi-Task Learning. CoRR, 2023, doi: 10.48550/arXiv.2310.09478.

[41] Peng, Z.—Wang, W.—Dong, L.—Hao, Y.—Huang, S.—Ma, S.—Wei, F.: Kosmos-2: Grounding Multimodal Large Language Models to the World. CoRR, 2023, doi: 10.48550/arXiv.2306.14824.

[42] Wang, W.—Chen, Z.—Chen, X.—Wu, J.—Zhu, X.—Zeng, G.—Luo, P.—Lu, T.—Zhou, J.—Qiao, Y.—Dai, J.: VisionLLM: Large Language Model Is Also an Open-Ended Decoder for Vision-Centric Tasks. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (Eds.): Advances in Neural Information Processing Systems 36 (NeurIPS 2023). Curran Associates, Inc., 2023, pp. 61501–61513, https://proceedings.neurips.cc/paper_files/paper/2023/file/c1f7b1ed763e9c75e4db74b49b76db5f-Paper-Conference.pdf.

[43] Bai, J.—Bai, S.—Yang, S.—Wang, S.—Tan, S.—Wang, P.—Lin, J.—Zhou, C.—Zhou, J.: Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. CoRR, 2023, doi: 10.48550/arXiv.2308.12966.

[44] Qiao, Y.—Qi, Y.—Hong, Y.—Yu, Z.—Wang, P.—Wu, Q.: HOP+: History-Enhanced and Order-Aware Pre-Training for Vision-and-Language Navigation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, 2023, No. 7, pp. 8524–8537, doi: 10.1109/TPAMI.2023.3234243.

[45] An, D.—Qi, Y.—Li, Y.—Huang, Y.—Wang, L.—Tan, T.—Shao, J.: BEVBert: Multimodal Map Pre-Training for Language-Guided Navigation. CoRR, 2022, doi: 10.48550/arXiv.2212.04385.

[46] Wang, Z.—Li, J.—Hong, Y.—Wang, Y.—Wu, Q.—Bansal, M.—Gould, S.—Tan, H.—Qiao, Y.: Scaling Data Generation in Vision-and-Language Navigation. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11975–11986, doi: 10.1109/ICCV51070.2023.01103.

[47] Wang, L.—He, Z.—Dang, R.—Shen, M.—Liu, C.—Chen, Q.: Vision-and-Language Navigation via Causal Learning. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13139–13150, doi: 10.1109/CVPR52733.2024.01248.

[48] Ramakrishnan, S. K.—Gokaslan, A.—Wijmans, E.—Maksymets, O.—Clegg, A.—Turner, J. M.—Undersander, E.—Galuba, W.—Westbury, A.—Chang, A. X.—Savva, M.—Zhao, Y.—Batra, D.: Habitat-Matterport 3D Dataset (HM3D): 1000 Large-Scale 3D Environments for Embodied AI. Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS 2021), Track Datasets and Benchmarks, 2021, doi: 10.48550/arXiv.2109.08238.

[49] Xia, F.—Zamir, A. R.—He, Z.—Sax, A.—Malik, J.—Savarese, S.: Gibson Env: Real-World Perception for Embodied Agents. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9068–9079, doi: 10.1109/CVPR.2018.00945.

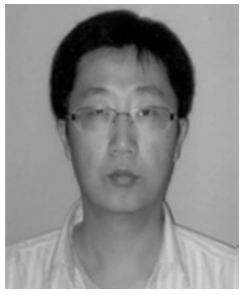
Haibo Li is currently pursuing a graduate degree with the Department of Computer Science and Technology, Xiamen University, Fujian, China. His research interests include computer vision, machine learning and rough set.

Zesheng Zhan is currently pursuing a graduate degree with the Department of Computer Science and Technology, Xiamen University, Fujian, China. His research interests include computer vision and machine learning.

Yaming Yang is currently pursuing a graduate degree with the Institute of Artificial Intelligence, Xiamen University, Fujian, China. His research interests include UAV intelligent agent and machine learning.

Fansheng Meng is currently serving as the General Manager of Taichang Technology (Hangzhou) Co., Ltd. With a long-term career in aircraft and computing-related fields, he has accumulated extensive experience in technology management and industrial innovation.

Yaoxin Chen is currently pursuing a graduate degree with the Department of Computer Science and Technology, Xiamen University, Fujian, China. His research interests include multimodal learning and machine learning.



Chong Zhao received his Ph.D. degree in the Department of Computer Science and Engineering from the Chinese University of Hong Kong. He is currently Assistant Professor in the Department of Computer Science, Xiamen University. His research interests include geometry processing, computer graphics and computer vision.

Qicong Wang received his Ph.D. degree in information and communication engineering from the Zhejiang University, Hangzhou, China. He is currently Associate Professor at the Department of Computer Science and Technology, Xiamen University, Xiamen, China. His research interests include computer vision, machine learning, big data analytics.